RELIABILITY FOR INTERCONNECT FABRICS

This is a continuation-in-part of U.S. Application No. 09/707,227, filed November 16, 2000, the contents of which are hereby incorporated by reference.

Field of the Invention:

The present invention relates to the field of networks. More particularly, this invention relates to reliability of networks.

Background of the Invention:

An interconnect fabric provides for communication among a set of nodes in a network. Communications originate within the network at a source node and terminate at a terminal node. Thus, a wide variety of networks may be viewed as a set of source nodes that communicate with a set of terminal nodes via an interconnect fabric. For example, a storage area network may be arranged as a set of computers as source nodes which are connected to a set of storage devices as terminal nodes via an interconnect fabric that includes communication links and devices such as hubs, routers, switches, etc. Devices such as hubs, routers, switches, etc., are hereinafter referred to as interconnect devices. Depending on the circumstances, a node may assume the role of source node with respect to some communications and of terminal node for other communications.

The communication requirements of an interconnect fabric may be characterized in terms of a set of flow requirements. A typical set of flow requirements specifies the required communication bandwidth from each source node to each terminal node. The design of an interconnect fabric usually involves selecting the appropriate arrangement of physical communication links and interconnect devices and related components that will meet the flow requirements.

An interconnect fabric that meets the minimum flow requirements under ideal conditions will not necessarily meet the flow requirements under other conditions, such as in the event of a failure of a communication link, interconnect device or related component. Therefore, network designers typically address these reliability considerations by building in excess capacity or redundancy to help meet flow requirements under adverse conditions. Prior techniques are largely ad hoc and, thus, tend to be time-consuming, error-prone and may result in an over-provisioned interconnect fabric.

Summary of the Invention:

A technique is disclosed for providing reliability to an interconnect fabric for communication among a set of nodes. The technique may be used to efficiently and programmatically produce a cost-effective interconnect fabric having a degree of reliability over a range of design problems.

In one aspect, a method provides reliability to an interconnect fabric for communication among a set of nodes. Ports associated with each node are partitioned into a first set of ports and a second set of ports. A first interconnect fabric is formed among the first set of ports for each node in response to a set of flow requirements. A second interconnect fabric is formed among the second set of ports.

In another aspect a system provides reliability to a design for an interconnect fabric for communication among a set of nodes. A set of design information includes a set of flow requirements for the interconnect fabric. A fabric design tool generates a first design for the interconnect fabric among of first set of ports for each node, the first design being in response to the flow requirements, and also generates a second design for the interconnect fabric among a second set of ports for each node.

The first interconnect fabric may be formed by generating arrangements of flow sets in response to a set of flow requirements, determining one or more port violations with respect to the first set of ports for each node and alleviating at least one of the port violations by merging a pair of the flow sets. The second interconnect fabric may be formed in response to the same set of flow requirements or in response to a relaxed set of flow requirements. Other features and advantages of the present invention will be apparent from the detailed description that follows.

Brief Description of the Drawings:

The present invention is described with respect to particular exemplary embodiments thereof and reference is accordingly made to the drawings in which:

Figure 1 shows a method for providing reliability to an interconnect fabric according to an embodiment of the present invention;

Figure 2 shows an arrangement of flow sets in an interconnect fabric for an example design according to an embodiment of the present invention;

Figure 3 shows how ports at each node may be partitioned into sets for the example design according to an embodiment of the present invention;

Figure 4 shows a method for forming interconnect fabrics among corresponding sets of ports according to an embodiment of the present invention;

Figures 5-6 show a first interconnect fabric for the example design evolving according to an embodiment of the present invention;

Figure 7-8 show a second interconnect fabric for the example design evolving according to an embodiment of the present invention;

Figure 9 shows first and second interconnect fabrics for the example design according to an embodiment of the present invention; and

Figure 10 shows a fabric design tool that may employ techniques of the present invention to provide reliability to an interconnect fabric design.

Detailed Description of a Preferred Embodiment:

Figure 1 shows a method 100 for providing reliability to an interconnect fabric according to an embodiment of the present invention. The method 100 partitions ports at each node into sets and forms interconnect fabrics among the ports of each set based on flow requirements among the nodes. Reliability is provided because multiple fabrics interconnect the nodes. In the event of a failure in one of the interconnect fabrics, another one of the interconnect fabrics may allow communications which would otherwise not occur due the failure.

In a step 102, a set of nodes to be interconnected by an interconnect fabric, and flow requirements among the nodes, are determined. Table 1 shows an example set of flow requirements for an interconnect fabric under design.

	Terminal Node 50	Terminal Node 52	Terminal Node 54
Source Node 40	a	b	c
Source Node 42	d	e	f
Source Node 44	--	g	h

The flow requirements in this example specify three source nodes (source nodes 40-44 in the figures below) and three terminal nodes (terminal nodes 50-54 in the figures below). If an interconnect fabric is to meet the flow requirements, it must contain communication paths between all pairs of the source and terminal nodes 40-44 and 50-54 having positive flow requirements and must have sufficient bandwidth to support all of the flow requirements simultaneously.

In one embodiment, the source nodes 40-44 are host computers and terminal nodes 50-52 are storage devices and the bandwidth values a-h are numbers expressed in units of megabits per second. Thus, the interconnect fabric under design may be storage area network.

In other embodiments, there may be multiple flow requirements between a given source and terminal node pair. In such embodiments, the cells of Table 1 would contain a list of two or more entries. And, depending on the circumstances, a node may assume the role of source node with respect to some communications and of terminal node for other communications.

Figure 2 shows an initial arrangement of flows for the flow requirements obtained at step 102 for this example. Each entry in the flow requirements table is represented by a communication path or flow between pairs of nodes. More particularly, flow a is between the source node 40 and terminal node 50, flow b is between source node 40 and terminal node 52, flow c is between source node 40 and terminal node 54, flow d is between source node 42 and terminal node 50, flow e is between source node 42 and terminal node 52, flow f is between source node 42 and terminal node 54, flow g is between source node 44 and terminal node 52, and flow h is between source node 44 and 54.

In addition, a desired level of reliability may be determined. For example, the desired level may be full-redundancy, in which the flow requirements continue to be met despite a failure of any single node port, link, or interconnect device in the interconnect fabric. As another example, the desired level may relaxed to something less than full-redundancy to provide a lower level of performance in the event of a failure. For example, to reduce costs, a lower level of bandwidth may be provided between pairs of nodes after a failure than would be desired under normal operating conditions. In one aspect, the bandwidth requirement for one or more flows could be reduced by a percentage or eliminated entirely.

At step 104, the ports of each node may be partitioned into sets. For example, the ports at each node may be divided into two sets. In other embodiments, the ports of each node could be further divided into an additional number of (k) sets. In which case, additional fabrics may used to interconnect the additional sets of ports to provide even greater redundancy and reliability. Figure 3 shows how ports at each node may be partitioned into two sets for the example design. In the example, each of nodes 40, 44, 50 and 52 has four ports. These ports may be partitioned into first and second sets, each with an equal number of ports. Also, in the example, node 42 has five ports. If a node has an odd number of ports (given by: 2n+1), they may be partitioned into two sets in which one set has one more port (given by: n+1) than the other set (given by: n). If a node has only one port, the port may be split among the sets by connecting an interconnect device having at least three ports, such as a hub or repeater, to the port. The remaining ports of the interconnect device may then be partitioned into the sets as ports belonging to the node.

In the example, a first set for the node 42 includes three ports while a second set includes two ports. And, in the example, node 54 includes two ports which may be partitioned into first and second sets of one each. The first set of ports for each node is shown in Figure 3 to the left of a dotted line which divides the node, while the second set of ports for each node is shown to the right of the dotted line.

In a step 106 (Figure 1), a first interconnect fabric is formed among a first set of ports for each node. For full redundancy, each of the sets of ports in the example has the same flow requirements. Thus, the flows of Figure 2 are shown in Figure 3 as being supported by the first set of ports for each node.

Figure 4 shows a method 200 for forming an interconnect fabric among sets of ports according to an embodiment of the present invention. The method 200 is disclosed in U.S. Application No. 09/707,227, filed November 16, 2000, the contents of which are hereby incorporated by reference, and may be performed during the step 106 of Figure 1. It will be apparent, however, that other techniques for forming an interconnect fabric, such as manual or other methods, may be used in the step 106.

The method 200 partitions the flow requirements of the interconnect fabric into flow sets and iteratively merges the flow sets while taking into account the feasibility and cost of the implementing the interconnect fabric.

At step 202, an arrangement of flow sets in the interconnect fabric is determined in response to the set of flow requirements for the source and terminal nodes. In one embodiment, step 202 is performed by generating a flow set for each flow specified in the flow requirements for the interconnect fabric. Thus, each of flows a, b, c, d, e, f, g and h of the example is initially included in a corresponding flow set having one flow.

At step 204, port violations which are associated with the arrangement of flow sets among the first set of ports are determined. In the example, port violations are determined for the first set of ports for each source node 40-42 and each terminal node 50-52. In general, the number of port violations is equal to the sum, over all flow sets, of the number of required physical communication links to the node from that flow set, minus the number of available ports in the first set of ports. Each flow set may require one or more physical communication links to a given source or terminal node in the network. In this example, the number of port violations for a node is equal to the number of flow sets connected to the node minus the number of available ports in first set of ports for the node because each flow set is carried by one physical communication link in the interconnect fabric.

In the example (Figure 3), the source node 40 has a port violation of one since each of its three flow sets requires one physical communication link to the source node 40 and the source node 40 has only two available ports in the first set. The source nodes 42-44 and the terminal node 50 have no port violations since the number of ports in the first set is equal to the number of flow sets. The terminal node 52 has a port violation of one and the terminal node 54 has a port violation of two.

In other examples, the number of available ports in the first set for the source nodes 40-42 and the terminal nodes 50-52 may differ and the number of physical communication links required by a flow set on a given source or terminal node it connects to may exceed one.

At step 206 (Figure 4), at least one of the port violations is alleviated by merging a pair of the flow sets. Step 206 initially involves selecting the pair of flow sets in the current interconnect fabric that are to be merged. Initially, a candidate pair of flow sets is chosen that would alleviate the port violation on a node with the greatest port violation if merged. If there is more than one such candidate pair then one of the candidate pairs that alleviates a port violation on a node having the next greatest port violation is chosen from among them. If there is more than one such candidate pair then a pair of them that would be least costly to merge is chosen. The cost of merging two candidate pairs may be determined by choosing the least expensive interconnect device that is feasible for the merged flow set.

In the current state of the example interconnect fabric shown in Figure 3, the terminal node 54 has a port violation of two, which is the worst port violation in the network, and its corresponding flow sets are candidates for merging at step 206. For example, the pair of flow sets having flows c and f or the pair of flow sets having flows f and h or the pair of flow sets having flows c and h may be chosen as candidate pairs. All of these candidate pairs, if merged, would alleviate one port violation from the terminal node 54 but none of them would alleviate any other port violations. Therefore, the cost of merging the candidate pairs may be used to select the candidate pair of flow sets at step 206. For example, the communication link and/or interconnect device and/or ports that are used to merge the flow sets having flows c and f may be less expensive than the equivalent components needed to merge the other equally qualified candidate pairs. For example, assuming that link cost is related to length, replacing two links over a longer distance with a single link would likely reduce costs more than if over a shorter distance.

The candidate pairs of flow sets considered at step 206 must be feasible to merge. An example of a pair of flow sets that is not feasible to merge is a pair for which an interconnect device of sufficient bandwidth is not available. For example, a flow set having 60 units of bandwidth cannot be merged with a flow set having 50 units of bandwidth if the highest bandwidth interconnect device available is 100 units. Another example of a pair of flow sets that is not feasible to merge is a pair that would exceed the available ports on every available interconnect device of the resulting flow set. Candidate pairs that are not feasible to merge are bypassed at step 206 in favor of other candidate pairs.

If port violations still exist in the interconnect fabric after step 206, then another candidate pair of flow sets is selected and merged in a repeat of step 206. The method 200 loops through steps 204-206 in an iterative fashion until all port violations are eliminated or until no further merges are feasible.

Figure 5 shows an interconnect fabric that results from the first pass through step 206. A flow set having an interconnect device 160, a flow of c, and a flow of f is the result of the merger of the flow set having the flow of c and the flow set having the flow of f. At this point, the interconnect fabric has a port violation of one at the source node 40 and a port violation of one at each of the terminal nodes 52 and 54. In the example, a next pass through step 206 may result in the selection and merger of the flow set corresponding to an interconnect device 162 and with the flows a and b which alleviates the port violation of the terminal node 50. Then, another pass through the set 206 may result in the selection and merger of the flow set corresponding to an interconnect device 164 and with the flows e and g which alleviates the port violation of the terminal node 54. A further pass through the step 206 may result in the selection and merger of the flow set corresponding to the interconnect device 160 and with the flow set including the flows c and f with the flow set including the flow h which alleviates the port violation of the terminal node 54.

Figure 6 shows an interconnect fabric that results from these additional passes through step 206. At this point, the interconnect fabric has no port violation remaining. Accordingly, Figure 6 shows a first interconnect fabric that interconnects the first set of ports for each node and that will support the flow requirements for the network. Note that if cost savings may be obtained by merging additional flow sets, such mergers may also be accomplished by making another pass through step 206.

Returning to the method 100 of Figure 1, once the first interconnect fabric has been formed in the step 106 among the first set of ports for each node, a second interconnect fabric may be formed in the step 108 among the second set of ports for each node. For the step 108, the method 200 or another method may be employed to form the second interconnect fabric. Assuming that the second interconnect fabric is to provide full-redundancy, the same flow requirements may be applied to the second interconnect fabric as were applied to the first fabric. A relaxed set of flow requirements may be used if a lower level of reliability is desired.

Figure 7 shows a second interconnect fabric for the example design. Assuming the method 200 is employed, the flow sets may be merged in the same manner as in Figure 6. Thus, in the example, interconnect devices 170, 172 and 174 connect to the second set of ports of each node, whereas, the devices 160, 162 and 164 connect to the first set of ports of each node. Figure 7 shows that the interconnect device 170 merges flows c, f and h and corresponds to the interconnect device 160 of Figure 6, the interconnect device 172 merges flows a and b and corresponds to the interconnect device 162 of Figure 6 and the interconnect device 174 merges flows e and g and corresponds to the interconnect device 164 of Figure 6.

However, because fewer ports are available in the second set at the node 42, there remains a port violation at node 42 for the second interconnect fabric. Thus, at least one additional merger is required. Note that in Figure 6, each flow set has at most one interconnect device associated with it and all of the flows for the flow set are routed through that interconnect device or a single communication link, if possible. Thus, the method 200 generally results in a fabric design of a single layer in which there are no links between device nodes.

Under certain circumstances, a single-layer fabric may not eliminate all of the port violations. In which case, the method 200, by itself, may not result in a fabric design in which there are no port violations. Returning to Figure 7, no additional merges of flow sets are feasible using the method 200. For example, to relieve the port violation, two of the flow sets having flows d, e or f would need to be merged. However, the flow set having flow e has already been merged by the device 174 and the flow set having flow f has already been merged by the device 170.

Thus, in one embodiment, the present invention may address remaining port violations by recursively generating one or more additional layers of interconnect fabric nodes. For port violations at source nodes, the problem (i.e. the current fabric configuration and the applicable design information) may be recast such that the device nodes are treated as the terminal nodes. Then, one or more additional layers of device nodes may be inserted between the source nodes and the device nodes to relieve the port violations at source nodes. This results in links between device nodes and, thus, increases the number of layers in the interconnect fabric. Similarly, for terminal port violations, the problem may be recast such that the device nodes are treated as the source nodes. Then, one or more additional layers of device nodes may be inserted in between the device nodes and the terminal nodes to relieve the terminal node port violations. This also results in links between the device nodes and, thus, increases the number of layers in the interconnect fabric. Such a technique is disclosed in co-pending U.S. Application No. _________, entitled, "Designing Interconnect Fabrics," and filed _________, the contents of which are hereby incorporated by reference and which is continuation-in-part of U.S. Application No. 09/707,227, filed November 16, 2000.

The above-technique may be performed during the steps 104 or 106 of Figure 1, as needed. Thus, in the example of Figure 7, because there remains a port violation at a source node 42, the devices 170, 172 and 174 may be recast as terminal nodes. In addition, while there is no device in the flow set having flow d, this link can itself be treated as a terminal node. Alternately, a "dummy" node that is equivalent to a two-port hub or repeater, may be inserted into the link and the dummy node treated as a terminal node. Then, the method 200 of Figure 4 may be applied by merging flow sets to alleviate the port violation.

Figure 8 shows the second interconnect fabric of the example with the addition of a device 176 which merges the flow set having flow d with the flow set having flow e. Note that there is now a link between the device 174 and the device 176 and that there is no longer a port violation at the node 42.

Accordingly, the addition of the device 176 adds a layer to the interconnect fabric.

Figure 9 shows first and second interconnect fabrics for the example design according to an embodiment of the present invention. As shown in Figure 9, the two fabrics simultaneously connect the nodes. Reliability is enhanced because, in the event of a failure of any single element of the first interconnect fabric, the flows among the nodes can still be achieved by the second interconnect fabric.

Figure 10 shows a system having a fabric design tool 300 that may employ the method 100 (and the method 200) to provide reliability to an interconnect fabric in response to a set of design information 330. The fabric design tool 300 may be implemented in software and/or hardware to perform its functions. The design information 330 in one embodiment includes a list of hosts (source nodes) and devices (terminal nodes) 310, a list of fabric node types 312, a list of link type data 314, a set of flow requirements data 316, a set of port availability data 318, a set of bandwidth data 320, and a set of cost data 322. The design information 330 may be implemented as an information store, such as a file or set of files or a database, etc.

The list of hosts and devices 310 may specify the hosts and devices which are to be interconnected by an interconnect fabric design 324.

The list of fabric node types 312 may specify available interconnect devices, such as hubs, routers, switches, etc.

The link type data 314 may specify a list of available communication links that may be employed in the interconnect fabric design 324 and any relevant constraints. There are numerous examples of available communication links including fiber optic links, fibre channel links, wire-based links, and links such as SCSI as well as wireless links.

The flow requirements data 316 may specify the desired flow requirements for the interconnect fabric design 322. The desired flow requirements may include bandwidth requirements for each pairing of the source and terminal nodes.

The port availability data 318 may specify the number of communication ports available on each source node and each terminal node and each available interconnect device.

The bandwidth data 320 may specify the bandwidth of each host and device port and each type of fabric node and link.

The cost data 322 may specify costs associated with the available communication links and interconnect devices that may be employed in the interconnect fabric design 324. The cost data 322 may also specify the costs of ports for source and terminal nodes and interconnect devices. Other relevant costs may also be indicated.

The interconnect fabric design 324 generated by the fabric design tool 100 includes a list of the physical communication links and interconnect devices and ports, etc. and may include cost data.

The foregoing detailed description of the present invention is provided for the purposes of illustration and is not intended to be exhaustive or to limit the invention to the precise embodiment disclosed. Accordingly, the scope of the present invention is defined by the appended claims.